ADVAY DEEPAK IYER

SOFTWARE / AI ENGINEER

advayiyer7@gmail.com · github.com/advayiyer7 · advayiyer.com

U.S. work authorized, no sponsorship required

EDUCATION

University of Southern California, Los Angeles, CA — B.S. Computer Engineering & Computer Science — May 2027

GPA: 3.94 cumulative (4.00 last 2 semesters) · **Viterbi Dean's List:** all 6 semesters

Coursework: Data Structures & Algorithms, Algorithms & Computing, Software Development, Operating Systems, Computer Architecture, Computer Systems Organization, Internetworking, Full-Stack Web Development, Embedded Systems, Distributed Systems & IoT, Digital Circuits

Certifications: Machine Learning Specialization (Stanford), Intro to Concurrent & Parallel Programming with GPUs (JHU), Google AI Essentials (Google)

EXPERIENCE

AI Engineer Intern — **Oraczen**, Los Angeles, CA (agentic-AI startup, enterprise supply chain & procurement) — May–Jul 2026

- Built a multi-tier procurement insight engine (deterministic spend detectors → LLM narrator → agentic explorer) that surfaced 16 of 17 analyst-validated insights from a $2.3B, 12K-row supplier dataset — vs 1 of 17 for a brute-force LLM baseline — at ~3× lower cost ($0.017 vs $0.053/run), keeping per-run cost flat as data scaled and every figure traceable to its exact source rows.
- Engineered a three-stage LLM confidence pipeline (Claude classifier → GPT-4o log-probability scorer → live web-search grounding) that labels a 1,000-SKU B2B catalog as branded/unbranded and auto-accepts 93% of items at a tunable confidence threshold, cutting human review to the 7% lowest-confidence tail with zero high-confidence errors at the strict operating point.
- Designed an agent-to-agent procurement protocol stack spanning identity, discovery, negotiation, settlement, and fulfillment layers, enabling autonomous buyer–supplier transactions across independent agents.

Software Engineer Intern — **Publicis Sapient**, Chicago, IL — Jun–Aug 2025

- Shipped a full-stack AI battery-recommendation product for a global automotive battery manufacturer as the sole intern on the project, owning the React / AWS Amplify front end and a Python / FastAPI service on EC2 end to end.
- Engineered the recommendation pipeline — primary database lookup with a web-scraping fallback for vehicles outside the client's compatibility data — and added a Memcached → ElastiCache caching layer that eliminated redundant, billable data-provider calls.

Undergraduate Researcher — **USC** (Prof. Bhaskar Krishnamachari) — Sep 2025–Present

- Building GNN-based schedulers for distributed-computing task graphs: training graph convolutional networks (GCNs) on the SAGA library to predict the optimal scheduling algorithm (HEFT and variants) per workload, benchmarked against adversarial problem generators and parametric schedulers.

Machine Learning Intern — **A3 Transforms India** (AI, automation & analytics consultancy) — May–Aug 2024

- Trained a logistic-regression graduation-risk model (Python, scikit-learn / JupyterLab) reaching 86% accuracy, surfacing the three most predictive factors via the model's learned coefficients.
- Built a multi-class logistic classifier recommending a student's optimal academic stream; pitched to and well-received by the Indian government.

Teaching Assistant (Calculus II) — **USC Dornsife** — Feb 2024–Present

- Facilitate 15 hrs/week of Supplemental Instruction, measurably improving student exam performance and engagement.

PROJECTS

Semantic LLM Cache — drop-in caching proxy for LLM APIs

Python · FastAPI · Redis (RedisVL) · Vector embeddings · Prometheus/ Grafana · Docker

- Built an OpenAI-compatible semantic caching proxy (drop-in via one base-URL change) that caches reworded LLM queries by meaning, cutting cache-hit latency from 5.7 s to ~100 ms (a ~98% reduction at the p95 tail) across a 2,000-request load test with 0 errors.
- Designed the proxy to route across multiple LLM providers (OpenAI, Anthropic, Ollama) with streaming responses, per-request control over match strictness and cache expiry, and de-duplication of concurrent identical requests to shield providers under load; monitored with live Prometheus/ Grafana dashboards.

Facet — live multimodal AI jewelry-design studio

Next.js · Bun · Gemini · Pgvector · Supabase

- Shipped Facet, a live multimodal AI jewelry-design studio that turns a user's prompts and reference images into generated designs and 2D→3D models.
- Architected an LLM-planner backend routing across three AI providers, with semantic search over each user's image library.

Tek — local-first desktop AI file agent

Electron · React · Python · LanceDB · ONNX

- Built Tek, a downloadable cross-platform desktop AI agent that searches your files fully offline with hybrid retrieval (vector + keyword) and a custom-trained cross-encoder reranker — 90% top-1 accuracy on a self-built 172-probe adversarial benchmark, +7.5 pts over the off-the-shelf baseline while shipping 4× smaller (23 MB int8 ONNX) and faster.
- Designed a 3-tier security architecture (sandboxed UI → Electron core as the only component allowed to touch files → local Python service) and packaged it to ship its own runtime — zero prerequisites, nothing leaves the machine.

SKILLS

Languages: Python, Java, C++, JavaScript, TypeScript, SQL

Web: React, Next.js, HTML/CSS (Tailwind), REST/JSON, FastAPI, Node.js/Express

Databases & Caching: PostgreSQL (pgAdmin), MySQL, MongoDB, Redis, Qdrant, LanceDB

Cloud & DevOps: AWS (EC2, Amplify, API Gateway, S3, IAM), Azure, Docker, Git/GitHub, Linux/bash, GitHub Actions (CI/CD), Postman

AI / ML: TensorFlow, PyTorch, Scikit-Learn, Neural Networks (CNNs, RNNs, GNNs), NLP (LLMs, RAG, Prompt Engineering), Hugging Face Transformers, LangChain, Reinforcement Learning, CUDA/GPU Programming